Het perfecte metadatamodel

DOOR MARKTONTWIKKELINGEN, VOLWASSEN BI-OMGEVINGEN EN VERANDERENDE INFORMATIEBEHOEFTEN, NEEMT DE BEHOEFTE AAN METADATA TOE. IN DE PRAKTIJK BLIJKT DE BESCHIKBARE METADATA VAAK NIET VOLDOENDE GESCHIKT OM AAN DE EISEN EN WENSEN VAN (WIJZIGING)BEHEER TE KUNNEN VOLDOEN. DAARNAAST MIST VAAK EEN INTEGRAAL METADATAMODEL WAARBIJ ALLE BENODIGDE METADATA IN ÉÉN CENTRALE REPOSITORY WORDT OPGESLAGEN. IN DIT ARTIKEL WORDT HET METADATAMODEL BESCHREVEN DAT INMIDDELS VOLWASSEN GENOEG IS OM AAN DE VEREISTEN VAN DIVERSE STAKEHOLDERS TE VOLDOEN.

Door: Jorik Kool MSc en Vincent Wylenzek MSc, Senior Consultants bij Ordina

Metadata, wat is het?

"Metadata is what you know and data is what you are looking for. Therefore, everything can function as metadata[1]". Deze quote van David Weinberger komt uit een heel ander veld dan BI, namelijk uit de bibliotheekwereld. Wij definiëren metadata als data over data en alle gegevens die nodig zijn om verwerking, autorisatie, transparantie, traceerbaarheid, wijzigingen, beschikbaar- en toegankelijkheid van de data te borgen. Daarnaast onderscheiden we operationele, technische en business metadata.

De metadatabehoefte neemt toe

Een succesvolle BI-oplossing voorziet in de huidige informatiebehoefte, is voorbereid op de te voorspellen informatiebehoefte en kan flexibel en snel aangepast worden aan toekomstige informatiebehoeften. Met de toenemende complexiteit van de markt, de grotere beschikbaarheid van data en sneller wijzigende businessmodellen is snel kunnen adapteren onmisbaar voor de business.

Tegenwoordig is BI verweven met de primaire processen en is er behoefte aan BI op operationeel, tactisch en strategisch niveau. Vooral het operationele niveau en de inbedding in het primaire proces stellen hoge eisen aan de tijdigheid en beschikbaarheid van BI. Waren voorheen gegevens met een leeftijd van een dag en beschikbaarheid tijdens kantooruren voldoende, nu moet BI 24/7 beschikbaar zijn en worden hybride architecturen geïmplementeerd waar real-time reporting direct op de bron of op een replica plaatsvindt en het 'traditionele' integratieproces nog via een DWH loopt.

Als laatste is de behoefte aan traceerbaarheid en reproduceerbaarheid van informatie ook enorm toegenomen. Denk bijvoorbeeld aan financiële instellingen en verzekeraars, deze moeten conform Basel of Solvency in staat zijn de herkomst van gegevens en mogelijke manipulatie aan te kunnen tonen.

Genoemde veranderingen resulteren in de volgende eisen:

- BI-systemen moeten snel en eenvoudig aangepast kunnen worden aan veranderende informatiebehoefte en /of een veranderende businessmodel (flexibiliteit)
- De informatie moet in hoge mate beschikbaar zijn (beschikbaarheid)
- Vrijwel iedereen in de organisatie moet bij deze gegevens kunnen (toegankelijkheid)
- Op het moment van creatie van een gegeven moet het beschikbaar zijn in het BI-systeem (tijdigheid)
- Alle bewerkingen en verplaatsingen van bron- tot en met BI-systeem worden opgeslagen (traceerbaarheid)
- Beheer en onderhoud van het BI-systeem moet zo efficiënt en gebruikersvriendelijk mogelijk zijn ingericht (beheerbaarheid)

Categorieën van metadata

In de inleiding werd metadata onderverdeeld in drie categorieën, namelijk in business, technische en operationele metadata. Hieronder volgt een toelichting van de diverse onderdelen per categorie:

BUSINESS METADATA

Begrippenlijst: Hier wordt de gezamenlijke begrippenlijst vastgelegd en bijgehouden via een grafische interface.

Business rules: De logica die toegepast wordt op de ingelezen data komt in veel verschijningsvormen. Het kan afleidingen of filters betreffen, maar ook prioriteiten. Zowel de business als ETL maken gebruik van de hier, centraal opgeslagen business rules.

Autorisatiematrix: Gevoelige data hoort slechts beschikbaar te zijn voor een beperkt gremium. Door het gebruik van een autorisatiematrix kan de organisatie een bewuste keuze maken welke informatieproducten en welke data binnen die informatieproducten voor wie inzichtelijk zijn. De inhoud van deze matrix wordt in deze entiteit opgeslagen en beheerd via een grafische interface.

Datavalidaties: Vaak wordt de data vanuit bronsystemen aangeboden volgens een vooraf afgesproken format. Voorbeelden van eenvoudige validaties zijn controles op datatype, veldlengte of aanwezigheid. Complexere varianten zijn controles op domeinen en onderlinge verbanden. De definities van deze validaties worden in deze entiteit vastgelegd.

TECHNISCHE METADATA

Groeiprognoses: Groeiprognoses zijn gebaseerd op statistieken over in gebruik zijnde database-objecten. Deze statistieken worden opgeslagen in de metadata repository en na elk laadproces bijgewerkt. Naast het monitoren van reguliere en verwachte groei heeft het bijhouden van groeiprognoses ook een nevendoel, namelijk het signaleren van onverwachte afwijkende patronen, wat kan duiden op een fout in het ETL-proces of een gewijzigde aanlevering.

Informatieproductenoverzicht: In het informatieproductenoverzicht wordt bijgehouden welke output het DWH genereert in de vorm van rapportages, kubussen, dashboards, feeds naar applicaties en overige bestandsformaten. Ook het gebruik van de informatieproducten wordt bijgehouden en er ligt een relatie met de autorisatiematrix.

Auditinformatie: De essentie van auditinformatie is aan kunnen tonen wat de oorsprong van een gegeven is en welke bewerkingen en business rules er op toegepast zijn tot en met het informatieproduct. De scope begint hier vaak bij binnenkomst van het DWH en eindigt bij het informatieproduct. Alle dataverplaatsingen en bewerkingen binnen de OLTP-omgeving vallen buiten de scope van de lineage, deze behoren tot de 'verantwoordelijkheid' van het bronsysteem. Binnen de auditinformatie worden entiteiten, attributen, ETL-mappings, business rules en datavalidaties aan elkaar gerelateerd. De ETL-mappingtabel staat centraal, omdat daarin de hele keten van bron tot en met informatieproduct wordt opgeslagen. Hieronder worden de auditcomponenten toegelicht:

Entiteiten: Alle entiteiten die onderdeel zijn van het DWH worden hier beschreven, zoals de definitie , de relatie met het bronsysteem waar hij deel van uitmaakt en specifieke eigenschappen van de entiteit (bijvoorbeeld statisch/volatiel, retentieperiode, versie et cetera). De entiteit is gerelateerd aan de begrippenlijst

Categorie	Onderdeel	Metadata-categorie
Flexibiliteit	Business rules	Business metadata
Beschikbaarheid	Geautomatiseerde verversing	Operationele metadata
	Groeiprognoses	Technische metadata
	Loginformatie	Operationele metadata
Toegankelijkheid	Autorisatiematrix	Business metadata
	Informatieproduct-enoverzicht	Technische metadata
Latency	Broninformatie	Operationele metadata
	Laadbereik	Operationele metadata
Traceerbaarheid (data lineage)	Business rules	Business metadata
	Datavalidaties	Business metadata
	Auditinformatie	Technische metadata
	Entiteiten	Technische metadata
	Attributen	Technische metadata
	ETL-mappings	Technische metadata
Beheerbaarheid	Versiebeheer (alle DDL, code, ETL-mappings en informatieproducten)	Technische metadata

Tabel 1. Vereisten – metadata-mapping

waar de functionele betekenis van de entiteit wordt beschreven.

Attributen: lle attributen die onderdeel zijn van de ontsloten entiteiten worden hier beschreven, zoals veldlengtes, datatypes, namen en definitie van de attributen. Het attribuut is gerelateerd aan de begrippenlijst waar de functionele betekenis van het attribuut wordt beschreven.

ETL-mappings: In de ETL-mapping-tabel wordt bijgehouden wat de bron- en doelattributen van de ETL-mappings zijn per laag en/of component in de logische BI-architectuur, zoals mappings tussen bron en staging, staging en DWH, DWH en data mart, data mart en kubus en tot slot kubus en rapport. Daarnaast wordt ook bijgehouden waar en wanneer business rules en datavalidaties worden toegepast.

Versiebeheer: Om binnen een bepaalde release, increment of iteratie alle wijzigingen bij elkaar te kunnen houden is het wenselijk één integrale versiebeheertool te gebruiken, waarbij alle onderdelen van het BI-project beheerd kunnen worden. Hierbij gaat het onder andere om functionele specificaties, datamodellen, ETL-software, informatieproducten en DDL.

OPERATIONELE METADATA

Broninformatie: In de beschrijving van een te ontsluiten bronsysteem hoort de methode en de frequentie van aanlevering of de beschikbaarheid voor uitlezen. Een indicatie of de bron actief is en wanneer de laatste verwerking op de bron heeft plaatsgevonden, maken het beeld compleet.

Laadbereik: Transactietabellen in de bron zijn vaak te groot om in z'n geheel over te halen. Als er een laadbereik is gespecificeerd voor een bron dan beperkt dat de dataset aanzienlijk. Voorwaarde is dat de transacties statisch zijn of dat er in de bron een wijzigingsdatum beschikbaar is voor filtering.

Geautomatiseerde verversing: Alle onderlinge afhankelijkheden van entiteiten en hun attributen zijn in kaart gebracht en de beschikbaarheid van bronsystemen is bekend. Dan is het inrichten van geautomatiseerde verversing een logische vervolgstap. Afhankelijk van de capaciteiten van de infrastructuur kan een plafond gezet worden op de mate van parallel verwerken. Op die manier vindt maximale uitnutting plaats van de beschikbare resources en blijft het tijdsraam van de verwerking zo kort mogelijk.

Loginformatie: Log informatie is informatie over statistieken en resultaten van laadprocessen, zoals bijvoorbeeld de geautomatiseerde of ad-hoc-verversing. Deze informatie kan bestaan uit een combinatie van native-tool-logging en custom-made logging.

ÉÉN CENTRALE METADATA REPOSITORY

Technische metadata wordt veelal binnen BI-tool-repositories bijgehouden. In de meeste organisaties wordt gebruik gemaakt van meerdere producten en leveranciers met allen hun eigen metadata repository (silo). Om een integraal beeld te krijgen is het belangrijk deze metadata te centraliseren in één repository. De tools blijven gebruik maken van hun eigen proprietary repositories, maar de stakeholders krijgen alleen toegang tot de metadata vanuit de centrale repository. Het heeft de voorkeur de repositories van de BI-tools slechts logisch te ontsluiten, bijvoorbeeld via views.

METADATA STAKEHOLDERS

Binnen een organisatie zijn er diverse stakeholders die al dan niet bewust gebruik maken van metadata. Technisch beheerders hebben behoefte aan operationele metadata. Functioneel beheerders richten zich op de business metadata. De releasemanager controleert de technische metadata. Ontwikkelaars en changemanagers focussen zich op een combinatie van de drie categorieën. Binnen het ETL-proces worden business rules toegepast uit de business metadata. De DBA is vooral geïnteresseerd in de operationele en technische metadata en de eindgebruikers zijn geïnteresseerd in business metadata. Vanwege de diversiteit aan gebruik is het belangrijk dat de metadata voor iedereen toegankelijk is ter inzage en aanvulling via een intuïtieve interface. Een goed voorbeeld van een tool die dit ondersteunt, is Master Data Services, beschikbaar in Microsoft SQL Server. Met een Excel add-in kan zowel technisch als functioneel beheer de inhoud van de metadata repository onderhouden zonder dat hiervoor specifieke technische kennis nodig is.

METADATA IN DE LOGISCHE ARCHITECTUUR

Als alle beschikbare metadata is opgeslagen in de metadata repository, kan deze via een data mart en eventueel een semantische laag naar de informatielaag ontsloten worden voor alle stakeholders Zo kan elke gebruikersgroep besluiten nemen op basis van relevante rapportages, hier acties op ondernemen en zo is de cirkel weer rond. Zie hieronder een schets van de logische architectuur van de centrale metadata repository.

AFSLUITING

Metadata is overal aanwezig en broodnodig; Het is hoog tijd het in zijn geheel toegankelijk en inzichtelijk te maken. Het helpt alle stakeholders om de voor hun relevante data te begrijpen en snel in te kunnen spelen op veranderende businessmodellen en inherente informatiebehoefte. Daarnaast bevordert het ook de toenadering tussen de verschillende groepen stakeholders, doordat de zwarte doos met complexiteit hierdoor transparanter wordt.

AUTEURS

Jorik Kool MSc en Vincent Wylenzek MSc zijn beiden Senior Consultant bij Ordina